TITLE: Knowledge Discovery in Databases
2-hour session.

To be presented at Avignon '94, May 30-June 4, 1994 by

Dr. Evangelos Simoudis
Research Scientist
Lockheed AI Center
0/96-20 B/254F
3251 Hanover Street
Palo Alto, CA 94304
simoudis@aic .lockheed.com
tel : (415) 354-5271
fax: (415) 424-3425

Dr. Usama Fayyad
Technical Group Leader
Artificial Intelligence Group
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, CA 91109
fayyad@aig. jpl.nasa.gov
tel : (818) 306-6197
fax: (818) 306-6912

Knowledge Discovery in Databases, or Database Mining, is the process of information extraction from very large databases. The extracted information is used either to explain phenomena in the mined data, or predict the outcomes based on new data. Database mining has recently received much attention, attested by the success of three Knowledge Discovery in Databases (KDD) AAM workshops, and a book, because organizations realize that while they generate data in ever-increasing rates much of it is never exploited. The rapid growth of data and information created a need and an opportunity for extracting knowledge from databases, and both researchers and application developers have been responding to that need. KDD applications have been developed for astronomy, biology, finance, insurance, marketing, medicine, and many other fields. Core Problems in KDD include representation issues, search complexity, the use of prior knowledge, and statistical inference. The unifying themes include the use of domain knowledge, managing uncertainty, interactive (human-oriented) presentation, and applications. The topics of interest covered in this presentation include:

Data preparation
% Data preprocessing techniques; filling missing values, data cleaning
% Selecting the data to be mined through database management, statistical analyses, and knowledge-based methods.
% Dimensionality reduction and feature space selection
Database mining techniques
% Artificial intelligence techniques (inductive learning and conceptual clustering).
% Neural networks (backpropagation, probabilistic neural network) .
% Case-based reasoning techniques (nearest neighbor and its variants).
% Deductive databases.
% Data visualization, and statistical analysis.
% Overview of Traditional approaches in Pattern Recognition

We will present the basic characteristics of each technique and illustrate each technique's applicability with examples from systems that have been presented in literature.

One of the hardest problems in database mining involves the selection the most appropriate technique for extracting information from a particular data set. This problem is more acute when several techniques need to be used cooperatively. We will present criteria for selecting each of the discussed techniques based on the characteristics of the data and the analyst's goals.

We will review four database mining systems developed by the authors for financial, manufacturing, Astronomy, and large image databases in palnetary sciences.